August 29th, 2017

1 Data Science Tool-set

Configuring a data science environment can be a pain. Dealing with inconsistent package versions, having to dive through obscure error messages, and having to wait hours for packages to compile can be frustrating. This makes it hard to get started with data science in the first place. Fortunately, Docker is here to the rescue!

Docker makes it fast and easy to create new data science environments, and use tools such as Jupyter notebooks along with tools like NumPy, Pandas, Matlibplot, etc. to explore your data. Docker containers are a layer over Linux containers that makes them easier to manage and distribute. Docker makes it easy to download images that correspond to a specific set of packages, and start them quickly. Docker is cross-platform, and works on Mac, Windows, and Linux.

In this tutorial, well cover the basics of Docker, how to install it, and how to leverage Docker containers to quickly get started with data science on your own machine using a docker data science container.

1.1 Getting started with Docker

The first step is installing Docker. Theres a graphical installer for Windows and Mac that makes this easy. Here are the instructions for each OS:

1. Windows : `https://docs.docker.com/docker-for-windows/`

2. MacOS `https://docs.docker.com/docker-for-mac/`

As part of this installation process, you'll need to use a shell prompt. The shell prompt, also called the terminal or the command line, is a way to run commands on your machine from a text interface instead of graphically.

On Windows, the shell is called Command Prompt, while on Mac and Linux it is called Terminal.

Youll need to use this same shell prompt whenever the rest of this post mentions having to run a Docker command or type a specific command.

1.2 Downloading the Image

Once Docker is downloaded, the next step is to download the image/configuration you want. Open the terminal window if using Mac or the command prompt if you are using Windows. Then type:

```
docker pull dataquestio/python2-starter
```

This pull download the image along with all the data science libraries.

1.2.1 Make a folder

Make a folder on your local machine that will correspond to where you want the notebooks/code stored. This folder will contain all of your work, and will persist on your local machine, even if you terminate the docker container. For example, I made a folder under the folder: /Users/msalloum/econ128_notebooks.

1.2.2 Running the image

Once you download the image, you can run it using docker run. We need to pass in a few options to ensure that its configured properly.

1. The -p flag sets the ports so that we can access the Jupyter notebook server from our machine.
2. The -d flag runs the container in detached mode, as a background process.
3. The -v flag lets us specify which directory on the local machine to store our notebooks in.

The full command looks like:

```
docker run -d -p 8888:8888 -v /Users/msalloum/econ128-notebook:/home/ds/notebooks dataquestio/python2-starter
```

You should change /Users/msalloum/econ128-notebook to whatever folder you created to store your notebooks in.

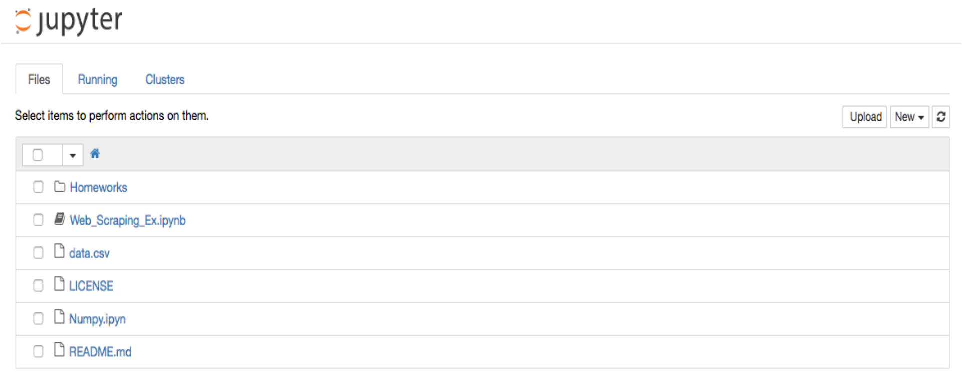
Executing docker run will create a Docker container. This is isolated from your local machine, and it may be helpful to think of it as a separate computer. Inside this container, Jupyter notebook will be running, and well be able to access many data science packages.

1.2.3 Viewing the notebook server

The next step is easy. simply open a web-browser and type :

```
localhost:8888
```

You should see the notebook running, and it will open the folder specified (for example, for me it is /Users/msalloum/econ128-notebook). A snapshot of my notebook is shown below; yours should may not contain files until you either place files or create new files in that folder.



Select items to perform actions on them.

□ Homeworks
□ Web_Scraping_Ex.ipynb
□ data.csv
□ LICENSE
□ Numpy.ipyn
□ README.md

Figure 1: Jupyter Notebook